

深度学习

姓名：罗飞彬

班级：信研 2305 班

学号：2023200821

指导教师：李瑞瑞

中文命名实体识别

摘要：针对目前中文命名实体识别研究中存在的语义特征提取不充分、不全面等问题，Transformers(BERT)在各种相关 NLP 任务中显示出惊人的改进，并且已经提出了连续的变体来进一步提高预训练语言模型的性能。在本文中，我们的目标是重新审视中文预训练语言模型，以检验它们在非英语语言中的有效性。本文基于 RoBERT 模型进行微调，实验结果表明，在许多 NLP 任务上表现良好。

关键词：命名实体识别；预训练模型；微调

ABSTRACT: In response to the current problems of inadequate and incomplete semantic feature extraction in Chinese named entity recognition research, Transformers (BERT) has shown striking improvements in a variety of related NLP tasks, and successive variants have been proposed to further improve the performance of pre-trained language models. In this paper, our goal is to revisit Chinese pre-trained language models to examine their effectiveness in non-English languages. This paper is based on the RoBERT model for fine-tuning, and experimental results show good performance on many NLP tasks.

KEYWORDS: Named entity recognition; pre-trained models; fine-tuning

1 引言

1.1 研究背景

命名实体识别 NER(named entity recognition)是指从一段自然语言文本中标注相关实体的位置和类型。例如新闻领域的人名、地名和机构名等名称的识别，医疗领域的疾病和症状等实体的识别。NER 通常利用序列标注方式联合识别实体边界和确定实体类型。NER 在知识图谱构建、信息抽取、信息检索、机器翻译、自动问答及舆情监测等任务中都有广泛应用，是自然语言处理的基础之一。在信息安全领域，利用 NER 分析相关实体，可以发现社交媒体中潜在的安全问题，也可以通过识别相关实体提供有效的信息来跟踪这些安全问题。

中文命名实体识别的相关背景源于自然语言处理领域的发展。随着互联网和社交媒体的普及，人们产生了大量的文本内容，处理这些文本数据成为一项重要的任务。其中，文本中包含的实体信息对于文本理解和信息提取至关重要，而命名实体就是文本中最具代表性的实体之一。

由于中文的语言特点和词汇特征，中文 NER 面临着一些独特的挑战。首先，中文词语常常由多个汉字构成，且没有空格进行分隔，因此需要对中文文本进行分词。其次，中文中的实体名称常常具有多种表达方式，如“北京市”、“北京”、“首都”等，需要进行同义词的匹配和处理。此外，还存在涉及多个实体的复杂命名实体识别问题。

为应对这些挑战，研究者们利用了深度学习等新技术来开发和优化中文 NER 模型，不断提高实体识别的准确性和效率。这些模型已经广泛应用于搜索引擎、信息提取、机器翻译等领域，并持续推动着自然语言处理技术的发展。

1.2 中文命名实体识别研究现状

对于中文 NER，最初的研究聚焦于专业名词的研究，张小衡[1]等根据机构名称的结构规律和形态标记等特点进一步总结规则，从 600 多万的三地语料库中识别高校名称实体，正确率达到了 97.3%。王宁等从专业名词识别的角度，充分考虑金融领域的特征，利用规则的方法专门针对公司名的识别问题进行了研究。该方法分析研究了金融新闻文本，总结了公司名的结构特征以及上下文信息，归纳形成知识库，并采取两次扫描的策略进行识别。在共 1336 篇真实金融新闻的数据集上进行实验，其中在封闭测试环境中的准确率和召回率分别为 97.13%和 89.13%，在开放测试环境中分别为 62.18%和 62.11%。

对于中文 NER，张华平[2]等借助 HMM 提出了基于角色标注的中国人名自动识别方法。该方法采取 HMM 对分词结果进行角色标注，通过对最佳角色序列的最大匹配来识别和分类命名实体，该方法解决了不具备明显特征的姓名的丢失、内部成词以及上下文成词的人名难召回的问题。俞鸿魁等提出一种基于层叠 HMM 的中文 NER 模型，该模型由三级 HMM 构成。在分词后低层的 HMM 识别

普通无嵌套的人名、地名和机构名等，高层的 HMM 识别嵌套的人名、地名和机构名。

对于中文 NER，李丽双[3]等提出一种基于 SVM 的中文地名的自动识别的方法，结合地名的特点信息作为向量的特征。此外，面对训练数据不足的难点，陈霄等针对中文组织机构名的识别任务，提出了一种基于 SVM 的分布递增式学习的方法，利用主动学习的策略对训练样本进行选择，逐步增加分类器训练样本的规模，进一步提高分类器的识别精度。

对于中文 NER，冯元勇[4]等在 CRF 框架中引入了小规模的常用尾字特征来降低特征集的规模，在提高模型训练速度同时保证识别准确率。燕杨等针对中文电子病历的 NER 问题，提出一种层叠 CRF，该模型在第二层中使用包含实体和词性等特征的特征集，对疾病名称和临床症状两类命名实体进行识别。与无自定义组合特征的层叠 CRF 相比，该模型的 F1 值提高了约 3 个百分点，和单层 CRF 相比，F1 值提高了约 7 个百分点。

半监督学习的 NER 方法主要采用自举的方法，该方法利用少量的标注数据进行训练，从而取得良好的实验结果。如 Teixeira[5]等提出一种基于 CRF 的自举训练方法，首先基于词典对 50000 条新闻标注人名，并使用标注好人名的数据作为训练集建立基于 CRF 的分类模型。然后使用 CRF 分类模型对初始种子语料库额外标注，并将其用于训练新的分类模型。该模型经过 7 次自举方法的迭代后，在 HAREM 数据集上进行实验表现良好。此外，Thenmalar[6]等不仅在英文语料中使用半监督的自举方法，还增加了泰米尔文语料进一步验证该方法的可行性。该方法利用少量训练数据中命名实体、单词和上下文特征来定义模式，分别对英文和泰米尔文进行 NER，两种语言的平均 F1 值为 75%。

对于中文 NER，针对结构复杂的产品名的识别任务，黄诗琳[7]等提出一种半监督学习方法，提取不同产品实体的结构特征和相互关系，构建一种三层半监督学习框架。首层结合规则和词典选取数据集中的候选数据；第二层利用相似度算法，把与种子集上下文相似的候选词加入正例中，这一步骤能解决数据稀疏问题；第三层是一个 CRF 的分类器用于识别相似度较低的实体。但因产品名的表达方式多样化，该方法与一般的 NER 方法相比，性能还存在一定的差距。在医学 NER 任务上，Long 等提出一个基于自举的 NER 方法，在自举训练过程中将命名实体特征集表示为类特征向量，候选命名实体的上下文信息表示为示例特征向量，这两种特征向量的相似程度决定了候选实体是否为命名实体。此外，针对少数民族语言的 NER 任务，王路路[8]等以 CRF 为基本框架，通过引入词法特征、词典特征以及基于词向量的无监督学习特征，对比不同特征对识别结果的影响，进而得到最优模型。

Etzioni[9]等提出了一个名为 KnowwitAll 的无监督 NER 系统，该系统以无监督和可扩展的方式自动地从网页中提取大量命名实体。Nadeau[10]等在 Etzioni 等的基础上进一步研究，该系统可以自动构建地名词典以及消解命名实体歧义，将构建的地名词典与常用的地名词典相结合。Han[11]等提出一个基于聚类主动学习的生物医学 NER 系统，该聚类方法通过使用底层分类器在文档中查找候选命名实体来进行聚类，因而更能反映命名实体的分布。无监督学习的 NER 方法既能解决有监督学习中需要大量带标注的训练数据的问题，也不需要少量标注的种子数据，但是这种方法需要提前确定聚类阈值并且性能较低，仍需进一步改善聚类方法。

对于中文 NER，2015 年 Wu[12]等利用卷积层生成由多个全局隐藏节点表示的

全局特征，然后利用局部特征和全局特征以识别临床文本中的命名实体。Wu 等提出了一种 CNN-LSTM-CRF，以获取短距离和长距离内容依赖，同时提出将 NER 和分词任务联合学习以挖掘这两个任务之间的内在联系，增强中文 NER 模型识别实体边界的能力，但该模型无法捕捉全局的上下文信息。因此，Kong 等提出一种融合多层次 CNN 和注意力机制的中文临床 NER 方法。该方法既能捕捉短距离和长距离的上下文信息，且注意力机制还能获取全局上下文信息，进一步解决了 LSTM 在句子较长时无法捕捉全局信息的问题。但该方法目前对稀有命名实体仍然存在难以识别的问题，因此，Gui[13]等将词典信息融合到 CNN 结构中，解决稀有实体识别的问题。可以发现，CNN 最大的特点是可以并行化，每个时间状态不受上一时间状态的影响，但其无法很好地提取序列信息。随着 RNN 的深入研究，CNN 和 RNN 常常混合使用。

Huang[14]等在 Collobert 等基础上，提出了多种基于 LSTM 的序列标注模型，包括 LSTM、Bi-LSTM 和 BiLSTM-CRF 等。首次将 Bi-LSTM-CRF 模型用于 NER，该模型不仅可以同时利用上下文的信息，而且可以使用句子作为输入。Gregoric[15]等在同一输入端采用多个独立的 Bi-LSTM 单元，通过使用模型间正则化来促进 LSTM 单元之间的多样性，能够减少模型的参数。Li 等提出一个模块化交互网络模型用于 NER，能同时利用段级信息和词级依赖。Xu 等提出一种有监督多头自注意网络的 NER 模型，利用自我注意力机制获取句子中词与词之间的关系，并引入一个多任务学习框架来捕捉实体边界检测和实体分类之间的依赖关系。

对于中文 NER，Zhang[16]等首次提出了基于混合字符和词典的 Lattice-LSTM 模型，通过门控单元，将词汇信息嵌入到每个字符中，从而利用上下文中有用的词汇提升 NER 效果。但是由于词汇的长度和数量无法确定，Lattice-LSTM 存在无法批量训练而导致模型训练较慢的问题。为了解决该问题，Liu 等提出了基于单词的 LSTM（WC-LSTM）。该方法在输入的向量中融入最优词汇的信息，在正向 LSTM 中融入基于该字开头的词汇信息，在反向 LSTM 中融入基于该字结尾的词汇信息。Ma 等也在 Lattice-LSTM 模型基础上做了改进，不修改 LSTM 的内部结构，只在输入层进行词与所有匹配到的词汇信息的融合，该方法还可以应用到不同的序列模型框架中，如 CNN 和 Transformer。

在中文领域，为了解决在 NER 过程中使用词典的最长匹配和最短匹配带来的问题，Ding[17]等提出了一种基于 GNN 并结合地名词典的 NER 方法，其目的使模型自动学习词典的特征。该模型首先根据地名词典构图，然后依次通过 GGNN 层、LSTM 层和 CRF 层进行实体的识别。Gui[18]等通过引入一个具有全局语义的基于词典的 GNN 模型来获取全局信息。此外，Tang[19]等进一步研究了如何将词汇信息整合到基于字符的方法中，提出一种基于单词-字符图卷积网络（WC-GCN），通过使用交叉 GCN 块同时处理两个有向无环图，并引入全局 GCN 块来学习全局上下文的节点表示。

对于中文 NER，Zhang[20]等利用远程监督的方法识别时间，提出了一种利用中文知识图谱和百度百科生成的数据集进行模型训练的方法，该方法不需要手动标注数据，且对不同类型的文本的适应性良好。此外，边俐菁[21]基于深度学习和远程监督的方法针对产品进行实体识别，利用爬虫整理得到的词典高质量地标注数据，按照词典完全匹配、完全匹配+规则、核心词汇+词性扩展+规则这三种方式进行实体识别，该方法能大大减少手工标注语料库的工作量。

基于 Transformer 方法典型代表是 BERT[22]类的预训练模型。Souza[23]等在 NER 任务上提出一种 BERT-CRF 模型，将 BERT 的传输能力与 CRF 的结构化预测相结合。

Naseem[24]等提出一种针对生物医学 NER 的预训练语言模型 BioALBERT，该模型在 ALBERT 中使用自我监督损失，能较好学习上下文相关的信息。Yang 等提出了一种分层的 Transformer 模型，应用于嵌套的 NER。实体表征学习结合了以自下而上和自上而下的方式聚集的相邻序列的上下文信息。

对于中文 NER，李妮[25]等提出了基于 BERTIDCNN-CRF 的中文 NER 模型，该模型通过 BERT 预训练模型得到字的上下文表示，再将字向量序列输入 IDCNN-CRF 模型中进行训练。Li 等为解决大规模标记的临床数据匮乏问题，在未标记的中国临床电子病历文本上利用 BERT 模型进行预训练，从而利用未标记的领域特定知识，同时将词典特征整合到模型中，利用汉字字根特征进一步提高模型的性能。Wu 等在 Li 等的基础上，提出了一个基于 RoBERTa 和字根特征的模型，使用 RoBERTa 学习医学特征，同时利用 Bi-LSTM 提取偏旁部首特征和 RoBERTa 学习到医学特征向量做拼接，解码层使用 CRF 进行标签解码。Yao 等针对制造文本进行细粒度实体识别，提出一种基于 ALBERT-AttBiLSTMCRF 和迁移学习的模型，使用更轻量级的预训练模型 ALBERT 对原始数据进行词嵌入，Bi-LSTM 提取词嵌入的特征并获取上下文的信息，解码层使用 CRF 进行标签解码。

2 研究概述

本文中，引入预训练模型进行 NER 任务训练，采用 RoBERT 模型，该模型由哈工大讯飞联合实验室开发。本文将该模型进行微调，并部署在服务器上。

2.1 预训练模型

BERT 全称为 Bidirectional Encoder Representation from Transformers（来自 Transformers 的双向编码表示），谷歌发表的发的论文 Pre-traning of Deep Bidirectional Transformers for Language Understanding 中提出的一个面向自然语言处理任务的无监督预训练语言模型。是近年来自然语言处理领域公认的里程碑模型。

BERT 有两部分：pre-training 和 fine-tuning。在 pre-training 阶段，会在没有标注数据且不同预训练任务上训练模型；在 fine-tuning 阶段，BERT 会根据预训练模型的参数初始化，然后在下游任务的标注数据进行 fine-tuned。

BERT 的结构如图一所示：

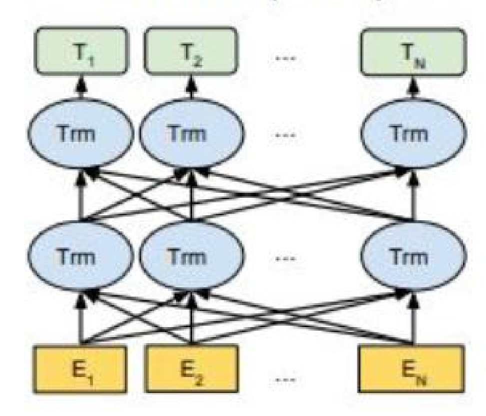


图 2-1 BERT 结构图

BERT 模型中使用的是 WordPiece embeddings，最后一层隐藏层的向量会作为每个 token 的表示。另外，有 3 个特殊字符如下：

- [CLS]：用于分类任务中每个序列的第一个 token。
- [SEP]：作为句子对（A，B）的分割符，句子首尾都有，具体可看输入输出表示部分。
- [MASK]：用于 masked ML 中 word 的替换。

RoBERTa：RoBERTa 模型是 BERT 的改进版(A Robustly Optimized BERT，即简单粗暴称为强力优化的 BERT 方法)。在模型规模、算力和数据上，与 BERT 相比主要有以下几点改进：

- 更大的模型参数量（论文提供的训练时间来看，模型使用 1024 块 V100 GPU 训练了 1 天的时间）
- 更大 bacth size。RoBERTa 在训练过程中使用了更大的 bacth size。尝试过从 256 到 8000 不等的 bacth size。
- 更多的训练数据（包括：CC-NEWS 等在内的 160GB 纯文本。而最初的 BERT 使用 16GB BookCorpus 数据集和英语维基百科进行训练）

另外，RoBERTa 在训练方法上有以下改进：

- 去掉下一句预测(NSP)任务
- 动态掩码。BERT 依赖随机掩码和预测 token。原版的 BERT 实现在数据预处理期间执行一次掩码，得到一个静态掩码。 而 RoBERTa 使用了动态掩码：每次向模型输入一个序列时都会生成新的掩码模式。这样，在大量数据不断输入的过程中，模型会逐渐适应不同的掩码策略，学习不同的语言表征。
- 文本编码。Byte-Pair Encoding（BPE）是字符级和词级别表征的混合，支持处理自然语言语料库中的众多常见词汇。原版的 BERT 实现使用字符级别的 BPE 词汇，大小为 30K，是在利用启发式分词规则对输入进行预处理之后学得的。Facebook 研究者没有采用这种方式，而是考虑用更大的 byte 级别 BPE 词汇表来训练 BERT，这一词汇表包含 50K 的 subword 单元，且没有对输入作任何额外的预处理或分词。

RoBERTa 建立在 BERT 的语言掩蔽策略的基础上，修改 BERT 中的关键超参数，包括删除 BERT 的下一个句子训练前目标，以及使用更大的 bacth size 和学习率进行训练。RoBERTa 也接受了比 BERT 多一个数量级的训练，时间更长。这使得 RoBERTa 表示能够比 BERT 更好地推广到下游任务。

2.2 微调（fine-tuning）

fine-tuning 的过程就是用训练好的参数（从已训练好的模型中获得）初始化自己的网络，然后用自己的数据接着训练，参数的调整方法与 from scratch 训练过程一样（梯度下降）。对于初始化过程，我们可以称自己的网络为目标网络，训练好的模型对应网络为源网络，要求目标网络待初始化的层要与源网络的层相同（层的名字、类型以及层的设置参数等等均相同）。

训练入口定义如图 2-2 所示

```
model.fine_tuneing(False)
train(10)

model.fine_tuneing(True)
train(10)
```

图 2-2

fine_tuneing 方法定义如图 2-3 所示

```
def fine_tuneing(self, tuneing):
    self.tuneing = tuneing
    if tuneing:
        for i in pretrained.parameters():
            i.requires_grad = True

        pretrained.train()
        self.pretrained = pretrained
    else:
        for i in pretrained.parameters():
            i.requires_grad_(False)

        pretrained.eval()
        self.pretrained = None
```

图 2-3

3 研究方法

3.1 模型训练

模型定义（下游模型）：

```
class Model(torch.nn.Module):
    def __init__(self):
        super().__init__()
        self.tuneing = False
        self.pretrained = None

        self.rnn = torch.nn.GRU(768, 768, batch_first=True)
        self.fc = torch.nn.Linear(768, 8)

    def forward(self, inputs):
        if self.tuneing:
            out = self.pretrained(**inputs).last_hidden_state
        else:
            with torch.no_grad():
                out = pretrained(**inputs).last_hidden_state

        out, _ = self.rnn(out)

        out = self.fc(out).softmax(dim=2)

        return out

    def fine_tuneing(self, tuneing):...
```
e

图 3-1

GRU 是 LSTM 网络的一种效果很好的变体，它较 LSTM 网络的结构更加简单，而且效果也很好，因此也是当前非常流形的一种网络。GRU 既然是 LSTM 的变体，因此也是可以解决 RNN 网络中的长依赖问题。

在 LSTM 中引入了三个门函数：输入门、遗忘门和输出门来控制输入值、记忆值和输出值。而在 GRU 模型中只有两个门：分别是更新门和重置门。具体结构如图 3-2 所示.

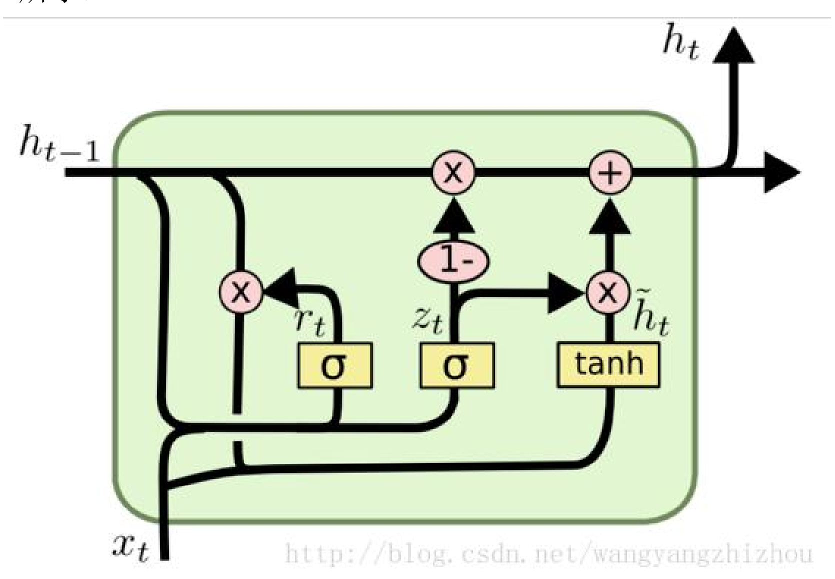


图 3-2

图 3-2 中的 Zt 和 Rt 分别表示更新门和重置门。更新门用于控制前一时刻的状态信息被带入到当前状态中的程度，更新门的值越大说明前一时刻的状态信息带入越多。重置门控制前一状态有多少信息被写入到当前的候选集 h~t 上，重置门越小，前一状态的信息被写入的越少。

3.2 系统模块

1.加载预训练模型，如图 3-3 所示。

```
pretrained = AutoModel.from_pretrained(model_path, from_tf=True)
```

图 3-3

2.加载分词器，如图 3-4 所示

```
tokenizer = AutoTokenizer.from_pretrained(model_path)
```

图 3-4

3.数据加载器，如图 3-5 所示

```
loader = torch.utils.data.DataLoader(dataset=dataset,
                                     batch_size=16,
                                     collate_fn=collate_fn,
                                     shuffle=True,
                                     drop_last=True)
```

图 3-5

4.学习率设置为，lr = 2e-5 if model.tuneing else 5e-4

5.优化器使用 AdamW

6.损失函数选用，torch.nn.CrossEntropyLoss()

7.训练过程如图 3-6 所示。

```
for epoch in range(epochs):
    for step, (inputs, labels) in enumerate(loader):
        if torch.cuda.is_available():
            inputs = inputs.cuda()
            labels = labels.cuda()

        # 模型计算
        # [b, lens] -> [b, lens, 8]
        outs = model(inputs)

        # 对outs和label变形,并且移除pad
        # outs -> [b, lens, 8] -> [c, 8]
        # labels -> [b, lens] -> [c]
        outs, labels = reshape_and_remove_pad(outs, labels,
                                              inputs['attention_mask'])

        # 梯度下降
        loss = criterion(outs, labels)
        loss.backward()
        optimizer.step()
        optimizer.zero_grad()
```

图 3-6

8.使用的数据集为，peoples_daily_ner

9.标签定义如图 3-7 所示。

```
"names": [
  "O",
  "B-PER",
  "I-PER",
  "B-ORG",
  "I-ORG",
  "B-LOC",
  "I-LOC"
],
```

图 3-7

10.对上述数据集进行拆分，如图 3-8 所示

```
"splits": {
  "train": {
    "name": "train",
    "num_bytes": 14972456,
    "num_examples": 20865,
    "dataset_name": "peoples_daily_ner"
  },
  "validation": {
    "name": "validation",
    "num_bytes": 1676741,
    "num_examples": 2319,
    "dataset_name": "peoples_daily_ner"
  },
  "test": {
    "name": "test",
    "num_bytes": 3346975,
    "num_examples": 4637,
    "dataset_name": "peoples_daily_ner"
  }
},
```


图 3-8

3.3 实验结果

1.部分训练数据，三列分别表示 loss、accuracy、去除”0”标签的 accuracy。结果表示如图 3-9 所示

```
1.509660005569458 0.7643610785463072 0.07373271889400922
1.4120501279830933 0.8619718309859155 0.14035087719298245
1.4009833335876465 0.8730337078651685 0.12403100775193798
1.3902822732925415 0.8945518453427065 0.34444444444444444
1.4019396305084229 0.8721088435374149 0.25396825396825395
1.3929541110992432 0.8814102564102564 0.28846153846153844
1.3965702056884766 0.8779527559055118 0.256
1.3743337392807007 0.8996960486322189 0.32653061224489793
1.4063763618469238 0.868020304568528 0.22962962962962963
1.372354507446289 0.9016602809706258 0.28703703703703703
```

图 3-9

用验证集数据进行预测并输出判断结果：并得出正确率和校正正确率（计算除了 0 以外元素的正确率,因为 0 太多了,包括的话,正确率很容易虚高）。训练结果如图 3-10 所示。

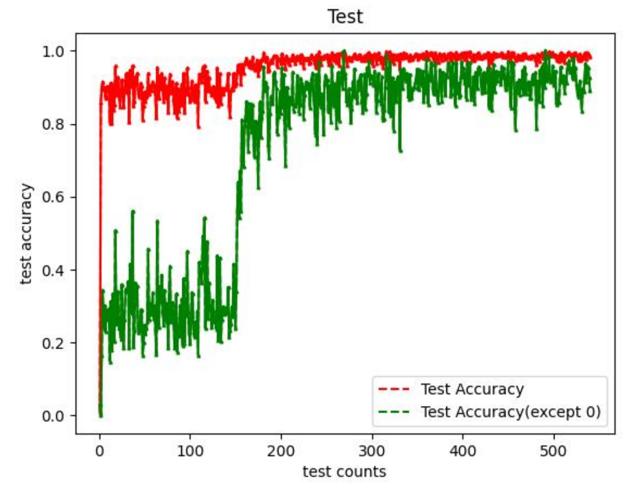


图 3-10

最终测试结果分别是 0.983961688040369（含 0 标签）、0.9080295790671218（不含 0 标签）

2.进行测试，测试代码如图 3-11 所示.

```
for step, (inputs, labels) in enumerate(loader_test):
    if step == 5:
        break
    print(step)

    with torch.no_grad():
        # [b, lens] -> [b, lens, 8] -> [b, lens]
        outs = model_load(inputs)

    # 对outs和label变形,并且移除pad
    # outs -> [b, lens, 8] -> [c, 8]
    # labels -> [b, lens] -> [c]
    outs, labels = reshape_and_remove_pad(outs, labels,
                                          inputs['attention_mask'])

    counts = get_correct_and_total_count(labels, outs)
    correct += counts[0]
    total += counts[1]
    correct_content += counts[2]
    total_content += counts[3]

print(correct / total, correct_content / total_content)
```

图 3-11

输出为：0.9893986121819583　0.9495287958115183

3.预测实验：

标签：['O', 'B-PER', 'I-PER', 'B-ORG', 'I-ORG', 'B-LOC', 'I-LOC']。

符号意义如图 3-12 所示

O = 单个词
PER = 人名
ORG = 机构
LOC = 地名

图 3-12

将标签用下标代替，其输出可看作如图 3-13 所示结果。



图 3-13

两种方式进行预测实验：

（1）用验证集数据进行预测并输出判断结果：

每个例子输出三行，第一行为输入，第二行为期望输出，第三行为模型实际输出。

==========================

[CLS]在这次采访中，长治人民医院宣教科郝彦南同志向我们提供了这样一串数据：

[SEP]
[CLS]7 • • • • • • • 长 3 治 4 人 4 民 4 医 4 院 4 宣 4 教 4 科 4 郝 1 彦 2 南 2 • • • • • • • • • • • • • • • [SEP]7
[CLS]7 • • • • • • • 长 3 治 4 人 4 民 4 医 4 院 4 宣 4 教 4 科 4 郝 1 彦 2 南 2 • • • • • • • • • • • • • • • [SEP]7
==========================
[CLS]据新华社莫斯科 4 月 28 日电俄罗斯总统叶利钦 28 日下午与基里延科总理会见后，任命了涅姆佐夫和赫里斯坚科两名副总理和第一批 7 名部长。[SEP]
[CLS]7 •新 3 华 4 社 4 莫 5 斯 6 科 6 •••••••俄 5 罗 6 斯 6 ••叶 1 利 2 钦 2 •••••••基 1 里 2 延 2 科 2 • • • • • • • • • 涅 1 姆 2 佐 2 夫 2 • 赫 1 里 2 斯 2 坚 2 科 2 • • • • • • • • • • • • • • [SEP]7
[CLS]7 •新 3 华 4 社 4 莫 5 斯 6 科 6 •••••••俄 5 罗 6 斯 6 ••叶 1 利 2 钦 2 •••••••基 1 里 2 延 2 科 2 • • • • • • • • • 涅 1 姆 2 佐 2 夫 2 • 赫 1 里 2 斯 2 坚 2 科 2 • • • • • • • • • • • • • • [SEP]7
==========================
[CLS]自治区党委常委、秘书长邱石元是工作中需要使用电话最多的人之一，但是他认为，只有做到长话短说，并做到移动电话[UNK]三不用[UNK]：在办公室不用、在会议室不用、回到家里不用，每月话费够用了。[SEP]
[CLS]7 自 3 治 4 区 4 党 4 委 4 • • • • • • 邱 1 石 2 元 2 ••• • • • • • [SEP]7
[CLS]7 自 3 治 4 区 4 党 4 委 4 • • • • • • 邱 1 石 2 元 2 ••• • • • • • [SEP]7
==========================
[CLS]前天，摩洛哥队外出训练时，所乘的崭新大巴坏了。[SEP]
[CLS]7 • • • 摩 3 洛 4 哥 4 队 4 • • • • • • • • • • • • • • • • [SEP]7
[CLS]7 • • • 摩 3 洛 4 哥 4 队 4 • • • • • • • • • • • • • • • • [SEP]7
==========================
[CLS]这其中，美国和俄罗斯队各具特色，可为代表。[SEP]
[CLS]7 • • • • 美 3 国 4 • 俄 3 罗 4 斯 4 队 4 • • • • • • • • • • [SEP]7
[CLS]7 • • • • 美 3 国 4 • 俄 3 罗 4 斯 4 队 4 • • • • • • • • • • [SEP]7
==========================
[CLS]该书是迄今为止国内较为全面、系统地研究周恩来经济思想的一部专著，有助于深化对周恩来思想的研究。[SEP]
[CLS]7 • • • • • • • • • • • • • • • • • • • 周 1 恩 2 来 2 • • • • • • • •••••••••周 1 恩 2 来 2 • • • • • • [SEP]7
[CLS]7 • • • • • • • • • • • • • • • • • • • 周 1 恩 2 来 2 • • • • • • • •••••••••周 1 恩 2 来 2 • • • • • • [SEP]7
==========================
[CLS]重头排列是嘉德这次拍卖的一大特点。[SEP]
[CLS]7 • • • • • 嘉 3 德 4 • • • • • • • • • • [SEP]7
[CLS]7 • • • • • 嘉 3 德 4 • • • • • • • • • • [SEP]7
==========================

[CLS]两个月后少女平静地离去，她的身边簇拥着俊平的朋友们，枕边还放着俊平为她捎去的书。[SEP]
[CLS]7•••••••••••••••••••俊 1 平 2••••••••••俊 1 平 2•••••••[SEP]7
[CLS]7•••••••••••••••••••俊 1 平 2••••••••••俊 1 平 2•••••••[SEP]7
==========================
[CLS]由他们创作演出的《征婚启示》开了话剧院团自己创作演出音乐剧的先河，受到广大观众的热烈欢迎。[SEP]
[CLS]7……………………………………………[SEP]7
[CLS]7 ••••••••••••••••话 3 剧 4 院 4 团 4 ••••••••••••••••••••••••[SEP]7
==========================
[CLS]此前，巴勒斯坦方面已同意接受美国的计划，并希望美国促使以色列也接受该计划。[SEP]
[CLS]7•••巴 5 勒 6 斯 6 坦 6•••••••美 5 国 6•••••••美 5 国 6••以 5 色 6 列 6•••••••[SEP]7
[CLS]7•••巴 5 勒 6 斯 6 坦 6•••••••美 5 国 6•••••••美 5 国 6••以 5 色 6 列 6•••••••[SEP]7
==========================
[CLS]这时，远处驶来一辆红色[UNK]菲亚特[UNK]，女主人下车问明缘由后，便拿出她车上的尼龙绳，想用她的车把我的车拖出来。[SEP]
[CLS]7 • • • • • • • • • • • • 菲 3 亚 4 特 4••••••••••••••••••••••••••••••••••••[SEP]7
[CLS]7 • • • • • • • • • • • • 菲 3 亚 4 特 4••••••••••••••••••••••••••••••••••••[SEP]7
==========================
[CLS]1986 年夏天，语文出版社安排吕老去怀柔一个招待所休息几天。[SEP]
[CLS]7••••••••语 3 文 4 出 4 版 4 社 4••吕 1••怀 5 柔 6•••••••••[SEP]7
[CLS]7••••••••语 3 文 4 出 4 版 4 社 4••吕 1••怀 5 柔 6•••••••••[SEP]7
==========================
[CLS]金人庆强调，全国税务系统要按照国务院的要求，加强征管，挖掘潜力，标本兼治，确保增收目标的实现；要顾全大局，为国分忧，把收入任务及时落实到基层；要抓紧清理漏管户，对重点税源加强专项检查，大力清理欠税，严格期初库存抵扣；要加强加油站、出租车的税收征管；要采取得力措施，认真落实调整商业企业增值税一般纳税人和交通运输企业抵扣增值税比例的税收政策；要强化税务稽查，进一步加快税务稽查队伍建设，充分发挥稽查职能，严厉打击偷逃税行为；要根据税源结构变化，及时调整征管力量，确保新的经济增长点同时也成为新的税收增长点。[SEP]
[CLS]7 金 1 人 2 庆 2 • • • • • • • • • • • • 国 3 务 4 院 4 •••
••
••
• •[SEP]7
[CLS]7 金 1 人 2 庆 2 • • • • • • • • • • • • 国 3 务 4 院

4 ••[SEP]7
==========================
（2）做了一个 QT 可视化界面，可呈现实验结果，如图 3-14 所示
、

in

共和党总统拜登当地时间12月6日恳求共和党人向乌克兰提供新的军事援助

out

共和党-ORG，拜登-PER，共和党-ORG，乌克兰-LOC

图 3-14

如图 3-14 中，输入“共和党总统拜登当地时间 12 月 6 日恳求共和党人向乌克兰提供新的军事援助。”输出为“共和党-ORG，拜登-PER，共和党-ORG，乌克兰-LOC”

4 总结

本文根据中文命名实体识别的课题研究背景和研究现状进行了分析和总结，对国内外研究现状进行了分析与总结。在文中介绍了所使用的预训练模型，并对其进行了微调，最后做了一个简单的可视化界面呈现实验结果，实现对中文的命名实体识别。

参考文献

[1]张小衡，王玲玲．中文机构名称的识别与分析[D]. , 1997.

[2]俞鸿魁，张华平，刘群，等．基于层叠隐马尔可夫模型的中文命名实体识别[J]．通信学报，2006, 2.

[3]李丽双，党延忠，廖文平，等. CRF 与规则相结合的中文地名识别[J]．大连理工大学学报, 2012, 52(2): 285-289.

[4]冯元勇，孙乐，张大鲲，等．基于小规模尾字特征的中文命名实体识别研究[J]．电子学报，2008, 36(9): 1833.

[5]Teixeira J, Sarmento L, Oliveira E. A bootstrapping approach for training a ner with conditional random fields[C]//Progress in Artificial Intelligence: 15th Portuguese Conference on Artificial Intelligence, EPIA 2011, Lisbon, Portugal, October 10-13, 2011. Proceedings 15. Springer Berlin Heidelberg, 2011: 664-678.

[6]Thenmalar S, Balaji J, Geetha T V. Semi-supervised bootstrapping approach for named entity recognition[J]. arXiv preprint arXiv:1511.06833, 2015.

[7]黄诗琳，郑小林，陈德人．针对产品命名实体识别的半监督学习方法[J]．北京邮电大学学报, 2013, 36(2): 20.

[8]古丽尼格尔·阿不都外力,吐尔根·依布拉音,卡哈尔江·阿比的热西提等.基于 Bi-LSTM-CRF 模型的维吾尔语词干提取的研究[J].中文信息学报,2019,33(08):60-66.

[9]Etzioni O、Cafarella M、Downey D 等人。Web-scale information extraction in knowitall：（初步结果）[C]//Proceedings of the 13th International Conference on World Wide Web.2004: 100-110.

[10]Nadeau D, Turney P D, Matwin S. Unsupervised named-entity recognition: Generating gazetteers and resolving ambiguity[C]//Advances in Artificial Intelligence: 19th Conference of the Canadian Society for Computational Studies of Intelligence, Canadian AI 2006, Québec City, Québec, Canada, June 7-9, 2006. Proceedings 19. Springer Berlin Heidelberg, 2006: 266-277.

[11]Yu M, Guo X, Yi J, et al. Diverse few-shot text classification with multiple metrics. In: Proc. of the NAACL-HLT. 2018. 1206−1215

[12]Wu F, Liu J, Wu C, et al. Neural Chinese named entity recognition via CNN-LSTM-CRF and joint training with word segmentation[C]//The World Wide Web Conference. 2019: 3342-3348.

[13]Gui, Tao, et al. "CNN-Based Chinese NER with Lexicon Rethinking." IJCAI. 2019.
[14]Huang Z, Xu W, Yu K. Bidirectional LSTM-CRF models for sequence tagging[J]. arXiv preprint arXiv:1508.01991, 2015.
[15]Žukov-Gregorič A, Bachrach Y, Coope S. Named entity recognition with parallel recurrent neural networks[C]//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). 2018: 69-74.
[16]Zhang Y, Yang J. Chinese NER using lattice LSTM[J]. arXiv preprint arXiv:1805.02023, 2018.
[17]Ding R, Xie P, Zhang X, et al. A neural multi-digraph model for Chinese NER with gazetteers[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. 2019: 1462-1467.
[18]Gui T, Zou Y, Zhang Q, et al. A lexicon-based graph neural network for Chinese NER[C]//Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP). 2019: 1040-1050.
[19]Tang Z, Wan B, Yang L. Word-character graph convolution network for chinese named entity recognition[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2020, 28: 1520-1532.
[20]Boudjellal N, Zhang H, Khan A, et al. Biomedical relation extraction using distant supervision[J]. Scientific Programming, 2020, 2020: 1-9.
[21]边俐菁.基于深度学习和远程监督的产品实体识别及其领域迁移研究[D].上海财经大学,2020.DOI:10.27296/d.cnki.gshcu.2020.000230
[22]Devlin J, Chang M W, Lee K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding[J]. arXiv preprint arXiv:1810.04805, 2018.
[23]Souza F, Nogueira R, Lotufo R. Portuguese named entity recognition using BERT-CRF[J]. arXiv preprint arXiv:1909.10649, 2019.
[24]Naseem U, Khushi M, Reddy V, et al. Bioalbert: A simple and effective pre-trained language model for biomedical named entity recognition[C]//2021 International Joint Conference on Neural Networks (IJCNN). IEEE, 2021: 1-7.
[25]李妮, 关焕梅, 杨飘, 等. 基于 BERT-IDCNN-CRF 的中文命名实体识别方法[J]. 《山东大学学报（理学版)》, 2020, 55(1): 102-109.